R-LiViT: A LiDAR-Visual-Thermal Dataset Enabling Vulnerable Road User Focused Roadside Perception

Jonas Mirlach[1*] Lei Wan[1,2*] Andreas Wiedholz[1*] Hannan Ejaz Keen[1] Andreas Eich[3]
[1]XITASO GmbH [2]Karlsruhe Institute of Technology [3]LiangDao GmbH
{jonas.mirlach, lei.wan, andreas.wiedholz, hannan.keen}@xitaso.com, andreas.eich@liangdao.de

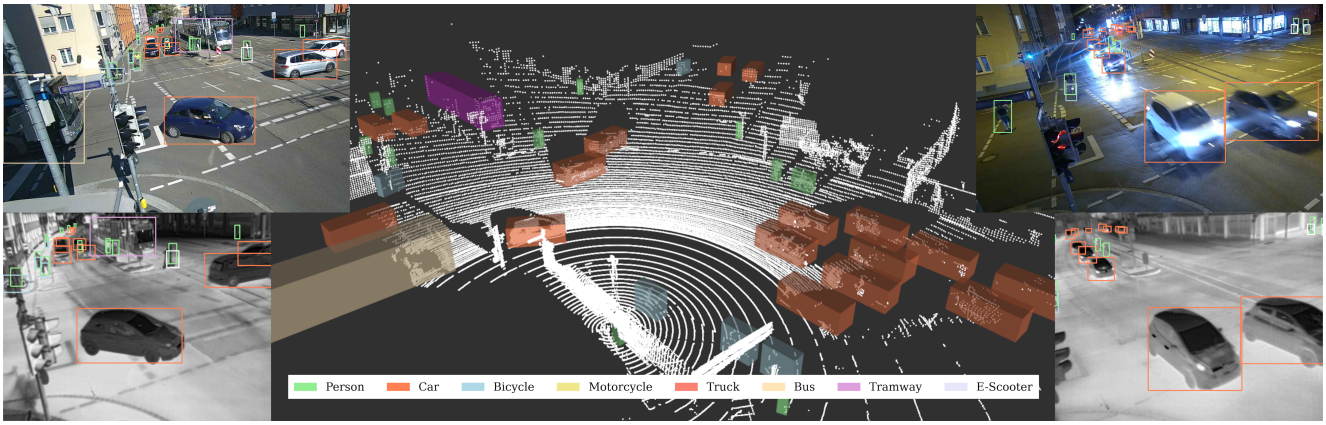


Figure 1. Illustrative visualization of an R-LiViT frame, with the right side depicting the same location at night.

Abstract

In autonomous driving, the integration of roadside perception systems is essential for overcoming occlusion challenges and enhancing the safety of Vulnerable Road Users (VRUs). While LiDAR and visual (RGB) sensors are commonly used, thermal imaging remains underrepresented in datasets, despite its acknowledged advantages for VRU detection in extreme lighting conditions. In this paper, we present R-LiViT, the first dataset to combine LiDAR, RGB, and thermal imaging from a roadside perspective, with a strong focus on VRUs. R-LiViT captures three intersections during both day and night, ensuring a diverse dataset. It includes 10,000 LiDAR frames and 2,400 temporally and spatially aligned RGB and thermal images across 150 traffic scenarios, with 7 and 8 annotated classes respectively, providing a comprehensive resource for tasks such as object detection and tracking. The dataset[1] and the code for reproducing our evaluation results[2] are made publicly available.

1. Introduction

In the domain of autonomous driving, research and development of early perception systems rely on large-scale datasets like KITTI [10] and nuScenes [4], which focus primarily on the vehicle perspective. However, in order to deal with occlusion, roadside perception systems play a critical role in achieving high-level autonomous driving and improving traffic management systems. By deploying roadside sensors, it is possible to detect road users more reliably and share this information with nearby vehicles to enhance the perception capabilities of connected and automated vehicles. Additionally, these systems can transmit data to traffic management platforms, enabling effective monitoring and control of traffic flow. The development of such systems requires high-quality annotated sensor data.

Although several roadside perception datasets have been published [2, 29, 33, 39], the majority predominantly emphasize vehicles while giving less attention to Vulnerable Road Users (VRUs), such as pedestrians and cyclists. Moreover, these datasets tend to focus primarily on LiDAR and visual (RGB) modalities. Recent research has demonstrated that, particularly for pedestrian detection, incorporating thermal imaging can be highly beneficial [12]. Ther-

*Equal contribution
[1]https://doi.org/10.5281/zenodo.16356714
[2]https://github.com/XITASO/r-livit

mal cameras capture heat radiation emitted by objects, making them effective for detection regardless of lighting conditions. This is particularly useful in both low-light environments and overexposed situations, such as glare. Consequently, the combination of RGB and thermal (RGB-T) has since been increasingly leveraged [3, 15, 16]. However, object detection and image fusion in RGB-T require temporally and spatially well-aligned data, which is currently available in only a limited number of datasets [12, 37]. While existing datasets include various combinations of LiDAR, RGB, and thermal modalities, few fully integrate all three, potentially limiting their effectiveness in ensuring comprehensive VRU safety.

To address these gaps, this paper introduces a multi-sensor system integrating LiDAR sensors, an RGB camera, and a thermal camera, creating R-LiViT (**R**oadside **Li**DAR-**Vi**sual-**T**hermal), the first multi-modal dataset incorporating these modalities from the roadside perspective focusing on urban traffic scenarios with significant VRU involvement. This setup allows comprehensive perception of all relevant road users across various lighting conditions. LiDAR provides precise 3D distance measurements, RGB cameras capture dense semantic information, and thermal cameras complement RGB by ensuring effective vision in low-light and overexposed situations. R-LiViT includes 10,000 LiDAR frames and 2,400 aligned RGB-T images across 150 traffic scenarios with annotations for 3D/2D object detection and 3D object tracking.

In summary, our contributions are:

- We present the first dataset for autonomous driving from a roadside perspective at intersections that integrates LiDAR, RGB, and thermal imaging, supporting object detection and tracking tasks.
- The dataset includes diverse traffic scenarios with a significant focus on VRUs, addressing the gaps in existing roadside perception datasets.
- With the standalone RGB-T subset, we introduce a challenging multiclass temporally and spatially aligned RGB and thermal dataset with high object and VRU density, contributing to the recent progress in RGB-T object detection and image fusion.

2. Related Work

This section provides an overview of related datasets. To the best of our knowledge, no existing LiDAR-RGB-T or standalone RGB-T dataset is specifically designed for object detection or tracking from a roadside perspective at intersections. Therefore, we compare our dataset with those that share some of its characteristics. First, we have a look at datasets that cover all three modalities LiDAR, RGB, and thermal. Second, we explore roadside perception datasets that adopt an infrastructure perspective. Third, we analyze RGB-T object detection datasets.

2.1. LiDAR-RGB-T Datasets

In the context of autonomous driving, there exist only a few datasets that incorporate the three modalities LiDAR, RGB, and thermal collectively, and they primarily focus on localization and depth estimation [5]. Brno Urban [21] includes these modalities but lacks proper annotations. ViViD++ [19] is a dataset that includes RGB, thermal, event, and depth cameras, capturing data under varying lighting conditions to support robust SLAM for autonomous navigation and robotics. More recently, Shin et al. [27] introduced a dataset that combines RGB, thermal, and LiDAR for multi-modal sensor fusion, primarily to improve depth estimation. KAIST Multi-Spectral [8] covers multiple tasks from a vehicle's perspective and is the only dataset more specifically designed for object detection, though it is not publicly available. Overall, existing LiDAR-RGB-T datasets provide little to no support for object detection and tracking in autonomous driving [5, 27]. R-LiViT fills this gap.

2.2. Roadside Perception Datasets

Roadside perception is vital to the perceptual accuracy of autonomous driving systems, as it addresses visual occlusions and enhances the perception capabilities of individual vehicles. To support research for roadside perception, several datasets have been released over the years.

In 2021, Yongqiang et al. [34] introduced the BAAI-VANJEE dataset, which features highway and urban intersection scenes captured under various weather and lighting conditions. DAIR-V2X-I [35], released in 2022 as part of the broader DAIR-V2X benchmark, provides infrastructure-side LiDAR and camera data from several intersections to support 3D object detection in cooperative settings. Wang et al. [29] developed IPS300+, an urban intersection recorded at a crossing near several universities, resulting in a relatively high presence of VRUs. The largest roadside perception dataset called LUMPI [2] contains data captured over different days in different weather conditions on a single intersection. Ye et al. [33] introduced Rope3D, a large-scale infrastructure-based dataset for monocular 3D object detection, with 3D annotations generated using LiDAR during data collection. In 2023, the TUMTraf intersection dataset [39], the successor to the A9 dataset [9], was published. It includes high-quality labels for object detection and tracking using point clouds and camera images, with a strong focus on vehicle detection. RCooper [13], released in 2024, enables infrastructure-to-infrastructure cooperative perception by capturing 3D-labeled LiDAR and camera data from multiple roadside units at both intersection and straight-road scenarios. Most recently, V2X-Real-I2I [31], as part of V2X-Real, was introduced, and provides multimodal data from paired infrastructure units covering a dense urban intersection, with a focus on benchmarking cooperative 3D detection fusion.

Dataset	Year	Modalities	# Frames	# Intersections	# Classes	Pedestrian density	Supported tasks
BAAI-VANJEE [34]	2021	RGB, LiDAR	2.5k PCL, 5k RGB	/	12	/	$OD_{2D\&3D}$
DAIR-V2X-I [35]	2022	RGB, LiDAR	10k	2	10	3.57	OD_{3D}
IPS300+ [29]	2022	RGB, LiDAR	14.1k	1	7	8.06	OD_{3D}, T
LUMPI [2]	2022	RGB, LiDAR	**90k PCL, 200k RGB**	1	6	8.39	OD_{3D}
Rope-3D [33]	2022	RGB, (LiDAR)	50k	1	**13**	3.42	$OD_{2D\&3D}$
TUMTraf [39]	2023	RGB, LiDAR	4.8k	1	10	0.77	OD_{3D}, T
RCooper [13]	2024	RGB, LiDAR	50k PCL, 30k RGB	1	10	0.45	OD_{3D}, T
V2X-Real-I2I [31]	2024	RGB, LiDAR	15k PCL, 31k RGB	1	10	**27.56**	OD_{3D}
R-LiViT (ours)	2025	**RGB, Thermal, LiDAR**	10k	**3**	7	9.61	**$OD_{2D\&3D}$, T**

PCL = Point Cloud, OD = Object Detection, T = Tracking, / = Information not accessible

Table 1. Comparison of R-LiViT and existing roadside perception datasets for autonomous driving. Pedestrian density refers to the annotated pedestrians per frame. All analyses are based solely on the publicly available versions of the datasets.

Dataset	Year	Perspective	Frequency	# Image pairs	# Classes	# Annotated objects	Object density	Person density
KAIST [15]	2015	Driving	20 FPS	**95,324**	1	**109,629**	1.15	1.15
FLIR-aligned [36]	2018	Driving	24 FPS	5,142	3	40,860	7.95	2.55
LLVIP [16]	2021	Surveillance	1 FPS	15,488	1	42,437	2.74	2.74
M^3FD [23]	2022	Multiplication	-	4,200	6	34,407	8.19	2.73
SMOD [6]	2024	Driving	2.5 FPS	8,676	4	32,874	3.97	2.14
R-LiViT (ours)	2025	Roadside	1.25 FPS	2,400	**8**	53,319	**22.22**	**8.76**

Table 2. Comparison of the RGB-T subset of R-LiViT and existing aligned RGB-T datasets. Object density refers to the annotated objects per image and person density to the annotated persons per image (for SMOD, we count additionally the rider class as person).

Table 1 provides an overview and comparison of the datasets discussed. While most of these datasets prioritize scale, they often lack diversity in terms of intersection types and primarily focus on vehicles with only a few datasets emphasizing the detection of VRUs [31, 34]. Our dataset specifically addresses this gap by ensuring a pronounced focus on VRUs, while covering three intersections to ensure a diverse dataset and supporting standard tasks such as object detection and tracking.

2.3. RGB-T Datasets

RGB-based perception degrades in low-light or night-time conditions, as well as in overexposed environments where intense illumination, such as headlight glare, can cause blending. To overcome this limitation, thermal cameras are increasingly being incorporated into perception systems, gaining substantial attention in recent years [3], especially for pedestrian detection [12]. Accordingly, several RGB-T datasets have been published in recent years. However, a key challenge for these datasets is ensuring proper temporal and spatial alignment of modalities, a requirement for most fusion algorithms that is met by only a few datasets [3, 12, 37].

The KAIST dataset [15], introduced in 2015, is one of the first and most widely used RGB-T datasets. It contains a large collection of paired RGB and thermal images with annotated pedestrians captured during both day and night. It is established as a benchmark for RGB-T pedestrian detection. The FLIR dataset was the first large publicly available RGB-T dataset with a broader range of object classes from the vehicle perspective and tailored specifically for autonomous driving. Due to spatial misalignment in the original version's images, Zhang et al. [36] created a revised, aligned version referred to as FLIR-aligned. LLVIP [16] from 2021 is a paired RGB-T dataset for pedestrian detection in low-light environments from a surveillance perspective. This dataset has also gained popularity as a benchmark for RGB-T pedestrian detection. The M^3FD dataset [23] features diverse scenarios, particularly focusing on environments where the thermal modality is expected to be valuable, such as adverse weather conditions. While not primarily designed for autonomous driving, it incorporates several traffic-related scenes. SMOD [6], published in 2024, is the most recent well-aligned RGB-T dataset designed for the driving perspective. It is specifically suited for autonomous driving use cases.

Table 2 provides an overview and comparison of the datasets discussed. While these datasets are of good quality, most of them have only limited classes annotated, focusing predominantly on pedestrians. Moreover, although LLVIP has a perspective similar to roadside, there are currently no dedicated RGB-T traffic datasets explicitly designed for roadside perception at intersections. Our standalone RGB-T dataset is the first high-quality dataset to address this gap.

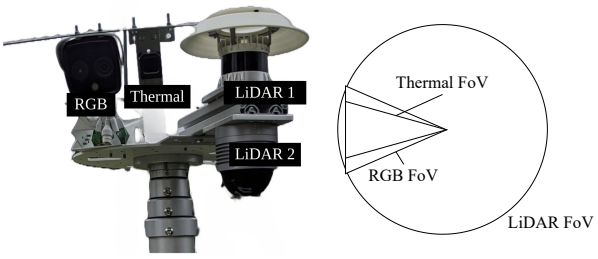
3. The R-LiViT Dataset

This section outlines in detail the creation and characteristics of the R-LiViT dataset.

3.1. Data Collection

Hardware

Our data collection setup includes one RGB camera, one thermal camera, and two LiDAR sensors. These sensors are mounted on top of a bar on a mobile sensor platform (see Figure 2a) that can be extended to a height of 4.5 meters. We use the Hikvision DS-2TD2628-3/QA with a resolution of 1280x720 as RGB camera and the FLIR ADK with a resolution of 640x512 as thermal camera. The RGB and thermal cameras record at 5 Hz and 60 Hz, respectively. The LiDAR setup consists of an Ouster OS1 BH with 64 beams and a RoboSense Bpearl with 32 beams. The second LiDAR sensor covers the blind spot of the first one, ensuring complete scene coverage by the LiDARs. The combined LiDAR system operates at 10 Hz.



(a) Sensor setup (b) Sensor FoVs

Figure 2. Sensor setup in the mobile sensor station and model of its field of views (FoVs).

Calibration and Alignment

The two LiDAR sensors are hardware-triggered, which leads to high precision in time synchronization (∼1ms) between them, generating the merged output mentioned above. The LiDARs, the RGB, and the thermal sensor are time synchronized with GPS. The LiDARs have integrated GPS modules and the RGB and thermal cameras rely on a GPS-NTP server integrated into the mobile sensor station. For the synchronization, we use the ApproximateTime module implemented in ROS with a maximum delay of 40 ms between data from the different sensors.

The system is calibrated as follows: the RGB camera is calibrated with the LiDAR sensors using an aluminum checkerboard (14x14 cm per tile and 6x8 tiles). Further, the thermal camera is calibrated with the RGB camera using a 5x7 circular calibration target with diagonal spacing of 95 mm and a circle diameter of 55 mm. For the thermal-to-RGB projection we estimate the homography using the calibration board resulting in a mean reprojection error of 0.038 Since the RGB images have a larger field of view (FoV) (see Figure 2b), we project the thermal image onto the RGB image during post-processing. To align their dimensions, the rest of the frame is padded. An example of this projection is shown in Figure 3.

Figure 3. Visualization of RGB-T projection and alignment: On the left is the undistorted RGB image, in the center the undistorted and projected thermal image, and on the right both are overlaid.

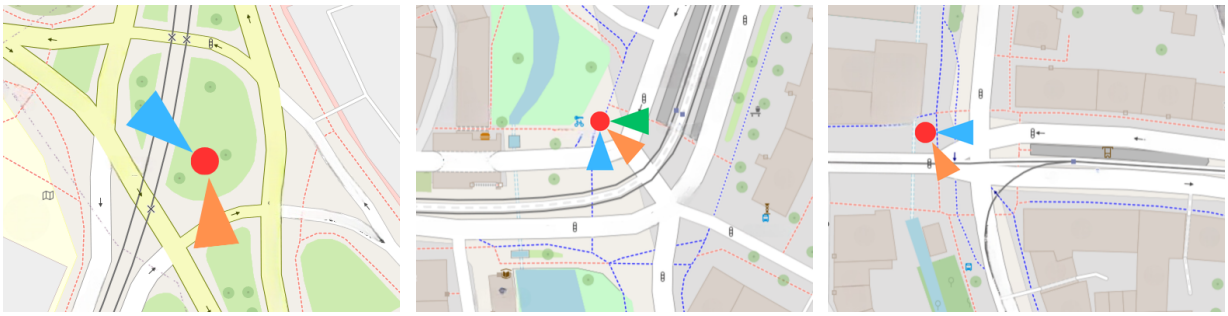
Data Recording and Selection

The dataset was collected in spring 2024 at three intersections (see Figure 4) in two German cities. These intersections were selected due to their high number of accidents involving VRUs, based on statistics from the Accident Atlas of the German Federal Statistical Office. This makes them particularly interesting for VRU detection in busy environments where high precision is essential.

As the FoVs of the cameras covers only a small part of the FoV of the LiDAR sensors, we rotated the mobile sensor station in order to cover all busy streets not only in the LiDAR but also in the camera data. This leads to a total of seven different views for the RGB and thermal recordings, visualized in Figure 4. The data was collected at daytime and at nighttime under clear weather conditions.

We manually recorded sequences of relevant and notable traffic scenarios, with human operators initiating and terminating the recordings based on their subjective evaluations of the street scene's activity level and VRU involvement. The data was captured at an aggregate frequency of 5 Hz. Each sequence was organized into frames of 50, corresponding to a duration of 10 seconds per sequence. In instances where the traffic situation remained of interest beyond this period, recording was continued, resulting in certain sequences within the dataset exceeding 10 seconds.

The final R-LiViT dataset comprises 200 sequences of 50 frames each. When merging consecutive and connected sequences, this number reduces to 150 independent traffic



(a) Intersection 1 (b) Intersection 2 (c) Intersection 3

Figure 4. Intersections where the data was collected. The red circle denotes the location of the mobile sensor station. The differently colored triangles indicate the different FoVs of the cameras after rotating the mobile sensor station.

scenarios of variable length. Overall, 65% of the dataset consists of daytime scenes and 35% of nighttime scenes. The daytime frames are distributed as 34%, 34%, and 32%, while the nighttime frames account for 29%, 24%, and 47% at intersections 1, 2, and 3, respectively.

Anonymization

To comply with the European General Data Protection Regulation (GDPR), we anonymized the RGB images. We manually detected and blurred all visible faces, license plates, and further markings that could potentially reveal a person's identity. In total, 5,122 faces and other objects were anonymized. Figure 5 depicts examples of anonymized regions. Since the images are captured from a height of 4.5 meters and traffic participants are mostly distant, the impact of anonymization on the images is limited. In Section 4.1, we demonstrate that the anonymization has no effect on the performance of common detection models.

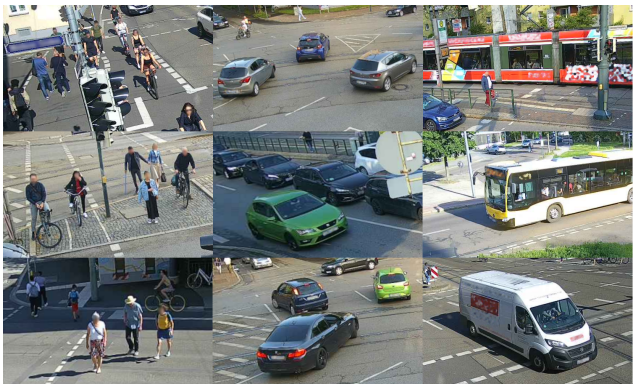
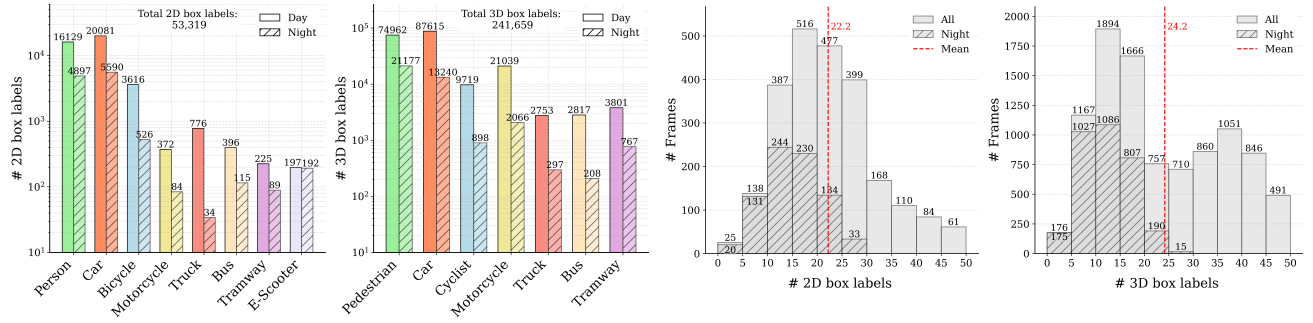


Figure 5. Cut-out examples of anonymized regions: In the left column anonymized pedestrians, in the center column anonymized license plates, and in the right column further anonymized markings such as advertisements.

3.2. Data Annotations

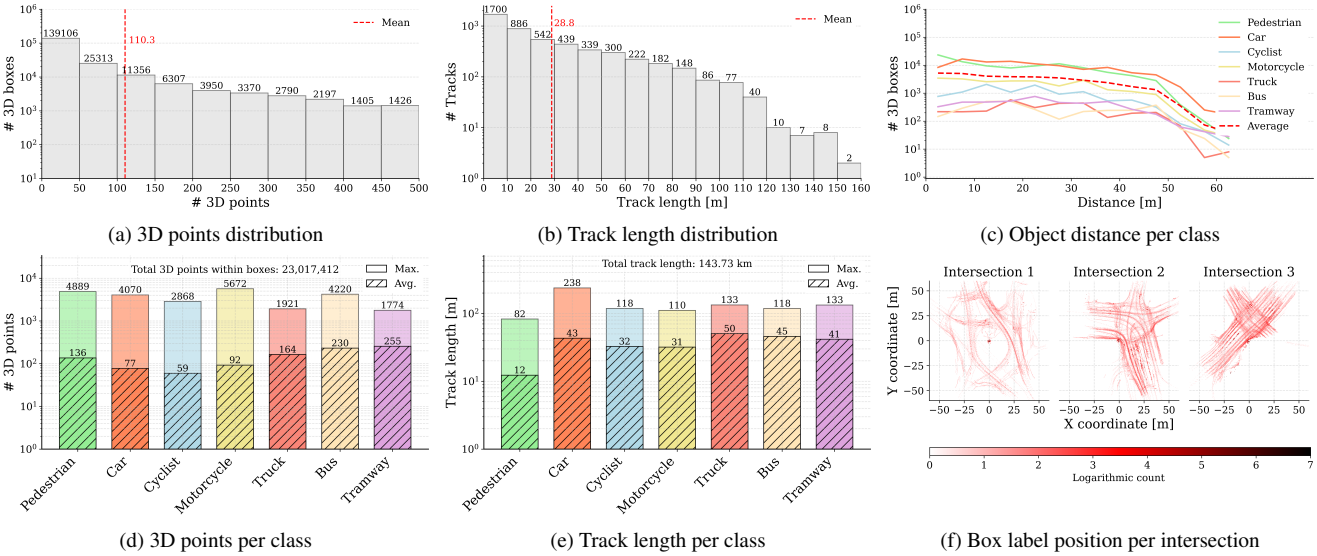
Data Labeling

The dataset includes two sets of bounding box annotations: One set of 2D bounding box annotations combined for the 2,400 RGB and thermal images and one set of 3D bounding box annotations for the 10,000 LiDAR frames. The labeled scenes are identical, with the only difference being that RGB data is annotated at a frequency of 1.25 FPS, while LiDAR data is annotated at 5 FPS. For the RGB-T images we labeled eight classes: *person*, *car*, *bicycle*, *motorcycle*, *truck*, *bus*, *tramway*, and *e-scooter*. For the LiDAR frames we labeled seven classes: *pedestrian*, *car*, *cyclist*, *motorcycle*, *truck*, *bus*, and *tramway*. We deliberately distinguish between *person* and *pedestrian*, as well as *bicycle* and *cyclist*, in the labels between the datasets, since in RGB-T annotations, a cyclist is labeled separately as a person and a bicycle, whereas in LiDAR annotations, a single bounding box represents the entire cyclist. The *e-scooter* class was not annotated in the LiDAR data because, on average, it does not contain enough points for reliable annotations. For the RGB-T, objects were annotated as long as they were detectable in any form. For the LiDAR, objects were annotated up to a distance of 50m to maintain reliability given the resolution of the point cloud. In addition to the box annotations, we provide tracking IDs for the LiDAR annotations. The labeling of RGB-T and LiDAR data was conducted independently. This approach was necessary because the RGB has a significantly deeper FoV into the distance, capturing more objects in that direction than LiDAR. Additionally, automatically projecting LiDAR annotations onto RGB-T introduces inaccuracies in the RGB-T labels, especially for smaller and more distant objects. By labeling them independently, we ensure higher annotation quality for the RGB-T data. For the labeling, a semi-automatic process was employed, with professional annotators manually reviewing and refining each frame. We predefine a train-test split, allocating 80% for training and 20% for testing. The



(a) 2D annotations per class (b) 3D annotations per class (c) 2D annotations distribution (d) 3D annotations distribution

Figure 6. Visualization of class distribution (note the logarithmic scale) and annotations per frame for both annotation sets.



(a) 3D points distribution (b) Track length distribution (c) Object distance per class

(d) 3D points per class (e) Track length per class (f) Box label position per intersection

Figure 7. Statistic of the 3D annotations (note the logarithmic scales). Plots inspired by [40].

test split is organized at the sequence level, ensuring that all images within the same sequence belong to the same split. The train-test ratio is also maintained consistently across all intersections, RGB perspectives, and times of the day.

Annotation Statistics

In total, there are 53,319 2D bounding boxes and 241,659 3D bounding boxes with the distribution of the classes shown in Figure 6a and Figure 6b. We can see here that the dataset contains a very high number of pedestrians and cyclists, i.e. VRUs. Figure 6c and Figure 6d display the distribution of label counts per frame. Naturally, the night images contain on average fewer objects per frame, as streets tend to be busier during the daytime.

The 2D bounding box annotations predominantly consist of small objects, accounting for 59.3%, followed by medium-sized objects at 30.3% and large objects at 10.4% (categorized according to the COCO [22] standard).

Figure 7 provides more insights into the 3D bounding box annotations. The 3D point distributions in Figure 7a and Figure 7d and the track length distributions in Figure 7b and Figure 7e generally appear natural and consistent, with no unexpected outliers. Remarkably, the overall average number of 3D points per box is relatively low. Figure 7c indicates that objects are distributed fairly evenly within the 50 meter radius of the sensor station. Finally, the heatmaps in Figure 7f visualize the tracks of the labeled objects in the point cloud at each of the three intersections.

3.3. Limitations

Similar to other datasets, our dataset creation approach is also subject to some limitations, which we transparently outline here so that users of the dataset are aware of them and can address them accordingly.

Since the data was collected using sensors that relied on software-based synchronization via GPS timestamps rather than hardware-based synchronization, the time alignment

between sensors is not always perfect. While synchronization between RGB and thermal is accurate, there are occasional slight temporal misalignments between RGB-T and LiDAR. This type of synchronization, which is prone to temporal misalignments, is also used in other datasets [29, 33]. Additionally, although the independent labeling approach enhanced the quality of the individual LiDAR and RGB-T datasets, it introduced minor inconsistencies in object annotations across the sets. This, coupled with differing FoVs, results in varying object coverage across the modalities. Another limitation is the lack of cross-modal tracking information. Tracking IDs were assigned only in the LiDAR domain, and no object ID mapping exists across datasets. We aim to address this in future iterations to enable tracking across all modalities. Finally, our setup includes a single RGB and thermal camera, which covers only a limited portion of the LiDAR FoV. Expanding to multiple RGB-T sensors in future work could further leverage the benefits of combining these three modalities for enhanced VRU detection.

4. Evaluation

To demonstrate the characteristics and applications of R-LiViT, we conduct a series of experiments and benchmarks.

For the benchmarking in the LiDAR domain, we employ three state-of-the-art 3D object detectors: the pillar-based detector PointPillars [18], the point-based detector PointRCNN [25], and the voxel-based detector PV-RCNN [26]. All models are trained using the OpenPCDet framework [28] and evaluated following the KITTI evaluation protocol [10]. Additionally, we assess three state-of-the-art 3D multi-object tracking (MOT) methods: AB3DMOT [30], Mahalanobis MOT [7], and SimpleTrack [20]. We use the official implementations of these trackers and employ detection results from PointPillars as their input to evaluate tracking performance, adhering to the AB3DMOT evaluation protocol [30]. All evaluations in the LiDAR domain are conducted per object category (car, cyclist, and pedestrian).

For the experiments in the RGB-T domain, we employ three established state-of-the-art single-modality object detectors: the two-stage detector Faster R-CNN [24], the one-stage detector YOLOv8, and the transformer-based RT-DETR [38]. For Faster R-CNN, we adopt the Torchvision implementation, incorporating a feature pyramid network and a ResNet-50 backbone. For YOLOv8 and RT-DETR, we use the Ultralytics [17] implementations with model sizes M and L, respectively. All detectors are pre-trained on the COCO dataset [22] and subsequently fine-tuned on R-LiViT. For evaluation, we follow the COCO protocol [22].

All experiments are conducted using three different random seeds, and results are averaged across them. The models are fine-tuned on the predefined training split and evaluated on the test split. To allow reproducibility, we publish the evaluation pipeline code[2].

4.1. Effect of Anonymization

Since we anonymized the RGB data to comply with GDPR, we want to demonstrate that this does not affect the performance of state-of-the-art object detectors. To evaluate this, we fine-tune the introduced models separately on the original and anonymized datasets using the full-resolution RGB images for training. The models are then evaluated on the test set using the original, non-anonymized images. The results, presented in Table 3, specifically examine the anonymized objects, namely persons and vehicles (car, truck, bus, and tramway). The findings indicate no significant difference in performance between models trained on anonymized versus non-anonymized data. This confirms the results also observed in similar experiments [1, 11, 14].

Model	Mode	AP_{per}	AP_{veh}
Faster R-CNN	original	37.65 ± 0.12	28.41 ± 0.81
	anonymized	37.55 ± 0.08	28.09 ± 0.61
YOLOv8	original	34.59 ± 0.17	34.92 ± 1.44
	anonymized	34.67 ± 0.20	35.25 ± 0.17
RT-DETR	original	31.88 ± 0.99	33.16 ± 0.82
	anonymized	31.29 ± 0.21	33.13 ± 1.46

Table 3. Effect of anonymization: AP_{per} denotes the mean average precision for persons, and AP_{veh} combined for cars, trucks, buses, and tramways. All values are reported with their sample standard deviation across the seeds.

4.2. Benchmarks

Here, we set a baseline for model performance on R-LiViT. First, we benchmark the introduced models on LiDAR 3D object detection, with the results presented in Table 4. Notably here, the point-based method PointRCNN performs substantially worse than other state-of-the-art approaches, primarily due to the dataset containing many objects with only a few points. As stated in Section 3.2, the average number of points within valid 3D bounding boxes is approximately 110, significantly lower than in other roadside perception datasets like TUMTraf [39]. This suggests that R-LiViT is well-suited for training models designed to handle scenarios with sparse point clouds per object.

Method	Car		Pedestrian		Cyclist	
	AP_{75}	AP_{50}	AP_{50}	AP_{25}	AP_{50}	AP_{25}
PointPillars	**46.73**	**58.38**	**23.07**	**30.64**	**30.11**	**37.58**
PointRCNN	28.26	31.58	4.62	5.20	11.76	13.09
PV-RCNN	46.72	56.19	19.79	26.85	28.43	33.88

Table 4. LiDAR object detection results on R-LiViT.

Method	Car			Pedestrian			Cyclist		
	sAMOTA ↑	AMOTP ↑	IDS ↓	sAMOTA ↑	AMOTP ↑	IDS ↓	sAMOTA ↑	AMOTP ↑	IDS ↓
AB3DMOT	**78.45**	**58.42**	21	**30.26**	**21.46**	3	**33.30**	**23.97**	**0**
SimpleTrack	64.40	48.96	**1**	18.80	15.20	**0**	25.15	18.18	**0**
Mahalanobis	61.53	47.37	2	19.99	18.27	2	22.01	16.44	2

Table 5. LiDAR multi-object tracking results on R-LiViT.

Model	Modality	Day							Night						
		AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AR^{100}_{per}	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AR^{100}_{per}
Faster R-CNN	RGB	35.06	61.32	36.51	18.73	43.31	58.77	43.44	24.98	48.76	23.73	19.75	32.62	56.09	38.30
	Thermal	19.05	40.60	15.24	9.23	23.26	44.44	27.91	15.21	39.18	8.59	14.69	17.10	32.59	25.55
	RGB+Thermal	31.59	53.95	33.19	16.39	41.18	58.64	46.74	20.47	41.93	17.24	17.90	27.82	45.60	44.67
YOLOv8	RGB	**44.87**	**63.02**	**49.85**	18.75	**48.72**	**73.74**	46.39	32.27	49.85	34.80	18.49	35.05	51.54	37.76
	Thermal	31.90	50.11	34.18	11.72	26.43	56.62	33.39	26.55	49.04	25.90	15.66	29.94	30.45	31.83
	RGB+Thermal	43.09	59.93	48.58	18.39	45.73	72.42	**50.42**	33.95	53.26	**37.40**	19.67	**37.34**	**53.53**	**45.34**
RT-DETR	RGB	37.50	59.67	40.57	18.74	37.59	57.79	40.68	32.67	54.60	35.27	20.16	27.91	45.75	34.15
	Thermal	29.08	50.70	30.12	13.47	24.36	47.45	31.84	27.03	50.48	25.42	18.88	27.48	26.53	28.74
	RGB+Thermal	37.12	58.78	40.84	**19.10**	36.00	58.21	44.93	**34.46**	**58.78**	36.57	**22.25**	30.81	45.24	41.14

Table 6. RGB-T object detection results on R-LiViT. AR^{100}_{per} denotes the average recall for persons considering up to 100 detections.

Second, we benchmark the introduced models on LiDAR 3D multi-object tracking, with the results presented in Table 5. AB3DMOT achieves the best performance across all three key categories. However, tracking performance on VRUs, such as pedestrians and cyclists, is significantly lower than on cars. This indicates that VRU tracking remains a challenging task, further complicated by the low average point cloud density in our dataset, making it more demanding compared to other benchmarks.

Third, we benchmark the introduced models on RGB-T 2D object detection, with the results presented in Table 6. For training and evaluation, we use a cropped frame of size 800x600 that fully covers both RGB and thermal FoVs to ensure a fair comparison. The thermal data is processed by replicating its single channel across three input channels of the models. For RGB+Thermal, we apply a simple late fusion technique, using non-maximum suppression with an IoU threshold of 0.8 to merge predictions from the RGB and thermal models. A key observation is that the models perform noticeably worse on our dataset compared to others like FLIR-aligned, LLVIP, or M^3FD (see reported results using similar models in [16, 23, 32]). This is primarily due to the high number of small objects, which remain challenging for state-of-the-art detectors. Furthermore, while RGB-based detection consistently achieves higher average precision compared to thermal, already primitive fusion techniques, such as the one applied here, can enhance overall performance, particularly in nighttime scenarios. More importantly, incorporating the thermal modality consistently improves recall for the person class, demonstrating its value in enhancing VRU perception. This underscores the potential of RGB-T fusion and motivates future research on optimizing fusion strategies for more robust detection systems.

5. Conclusion and Future Work

In this paper, we present R-LiViT, the first dataset for autonomous driving from a roadside perspective at intersections featuring LiDAR, RGB, and thermal modalities with a particular focus on VRUs. The data was collected from three different intersections in two cities, covering both day and night conditions, creating a diverse dataset. R-LiViT supports object detection and tracking but is also suitable for further tasks such as motion prediction and image-to-image translation. Compared to other datasets, R-LiViT offers a well-balanced class distribution especially having a very high density of pedestrians. Using this dataset can improve the accuracy of perception systems specifically for VRUs and therefore ensuring higher VRU safety. Additionally, we provide a benchmark of the performance of state-of-the-art object detection and tracking models on our dataset, setting a baseline for future work.

In future work, we plan to expand our setup and gather data across various weather conditions on a larger scale, and incorporate multiple perspectives, including the vehicle's perspective. This will facilitate further research on collaborative perception systems at complex intersections.

Acknowledgements

This work was supported by the German Federal Ministry for Economic Affairs and Climate Action (BMWK) within the program "Novel Vehicle and System Technologies" and the project "Valid Innovative Comprehensive Sensor System for Cooperative Automated Driving" (VALISENS), funding code 19A22009E.

References

[1] Mina Alibeigi, William Ljungbergh, Adam Tonderski, Georg Hess, Adam Lilja, Carl Lindström, Daria Motorniuk, Junsheng Fu, Jenny Widahl, and Christoffer Petersson. Zenseact open dataset: A large-scale and diverse multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20178–20188, 2023. 7

[2] Steffen Busch, Christian Koetsier, Jeldrik Axmann, and Claus Brenner. Lumpi: The leibniz university multi-perspective intersection dataset. In *2022 IEEE Intelligent Vehicles Symposium (IV)*, pages 1127–1134, 2022. 1, 2, 3

[3] Nicolas Bustos, Mehrsa Mashhadi, Susana K. Lai-Yuen, Sudeep Sarkar, and Tapas K. Das. A systematic literature review on object detection using near infrared and thermal images. *Neurocomputing*, 560:126804, 2023. 2, 3

[4] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[5] Spencer Carmichael, Austin Buchan, Mani Ramanagopal, Radhika Ravi, Ram Vasudevan, and Katherine A Skinner. Dataset and benchmark: Novel sensors for autonomous vehicle perception. *The International Journal of Robotics Research*, page 02783649241273554, 2024. 2

[6] Zizhao Chen, Yeqiang Qian, xiaoxiao Yang, Chunxiang Wang, and Ming Yang. Amfd: Distillation via adaptive multimodal fusion for multispectral pedestrian detection. *arXiv preprint arXiv:2405.12944*, 2024. 3

[7] Hsu-kuang Chiu, Jie Li, Rares Ambrus, and Jeannette Bohg. Probabilistic 3d multi-modal, multi-object tracking for autonomous driving. *IEEE International Conference on Robotics and Automation (ICRA)*, 2021. 7

[8] Yukyung Choi, Namil Kim, Soonmin Hwang, Kibaek Park, Jae Shin Yoon, Kyounghwan An, and In So Kweon. Kaist multi-spectral day/night data set for autonomous and assisted driving. *IEEE Transactions on Intelligent Transportation Systems*, 19(3):934–948, 2018. 2

[9] Christian Creß, Walter Zimmer, Leah Strand, Maximilian Fortkord, Siyi Dai, Venkatnarayanan Lakshminarasimhan, and Alois Knoll. A9-dataset: Multi-sensor infrastructure-based dataset for mobility research. In *2022 IEEE Intelligent Vehicles Symposium (IV)*, pages 965–970. IEEE, 2022. 2

[10] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012. 1, 7

[11] Jakob Geyer, Yohannes Kassahun, Mentar Mahmudi, Xavier Ricou, Rupesh Durgesh, Andrew S. Chung, Lorenz Hauswald, Viet Hoang Pham, Maximilian Mühlegg, Sebastian Dorn, Tiffany Fernandez, Martin Jänicke, Sudesh Mirashi, Chiragkumar Savani, Martin Sturm, Oleksandr Vorobiov, Martin Oelker, Sebastian Garreis, and Peter Schuberth. A2d2: Audi autonomous driving dataset, 2020. 7

[12] Bahareh Ghari, Ali Tourani, Asadollah Shahbahrami, and Georgi Gaydadjiev. Pedestrian detection in low-light conditions: A comprehensive survey. *Image and Vision Computing*, 148:105106, 2024. 1, 2, 3

[13] Ruiyang Hao, Siqi Fan, Yingru Dai, Zhenlin Zhang, Chenxi Li, Yuntian Wang, Haibao Yu, Wenxian Yang, Jirui Yuan, and Zaiqing Nie. Rcooper: A real-world large-scale dataset for roadside cooperative perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22347–22357, 2024. 2, 3

[14] Håkon Hukkelås and Frank Lindseth. Does image anonymization impact computer vision training? In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 140–150, 2023. 7

[15] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1037–1045, 2015. 2, 3

[16] Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. Llvip: A visible-infrared paired dataset for low-light vision. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 3489–3497, 2021. 2, 3, 8

[17] Glenn Jocher, Jing Qiu, and Ayush Chaurasia. Ultralytics YOLO, 2023. 7

[18] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12697–12705, 2019. 7

[19] Alex Junho Lee, Younggun Cho, Young-sik Shin, Ayoung Kim, and Hyun Myung. Vivid++ : Vision for visibility dataset. *IEEE Robotics and Automation Letters*, 7(3):6282–6289, 2022. 2

[20] Jiaxin Li, Yan Ding, Hua-Liang Wei, Yutong Zhang, and Wenxiang Lin. Simpletrack: Rethinking and improving the jde approach for multi-object tracking. *Sensors*, 22(15): 5863, 2022. 7

[21] Adam Ligocki, Ales Jelinek, and Ludek Zalud. Brno urban dataset - the new data for self-driving agents and mapping tasks. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3284–3290, 2020. 2

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. 6, 7

[23] Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5792–5801, 2022. 3, 8

[24] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2015. 7

[25] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3d object proposal generation and detection from point cloud. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 770–779, 2019. 7

[26] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10529–10538, 2020. 7

[27] Ukcheol Shin, Jinsun Park, and In So Kweon. Deep depth estimation from thermal image. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1043–1053, 2023. 2

[28] OpenPCDet Development Team. Openpcdet: An open-source toolbox for 3d object detection from point clouds. https://github.com/open-mmlab/OpenPCDet, 2020. 7

[29] Huanan Wang, Xinyu Zhang, Zhiwei Li, Jun Li, Kun Wang, Zhu Lei, and Ren Haibing. Ips300+: a challenging multi-modal data sets for intersection perception system. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2539–2545, 2022. 1, 2, 3, 7

[30] Xinshuo Weng, Jianren Wang, David Held, and Kris Kitani. 3D Multi-Object Tracking: A Baseline and New Evaluation Metrics. *IROS*, 2020. 7

[31] Hao Xiang, Zhaoliang Zheng, Xin Xia, Runsheng Xu, Letian Gao, Zewei Zhou, Xu Han, Xinkai Ji, Mingxi Li, Zonglin Meng, et al. V2x-real: a largs-scale dataset for vehicle-to-everything cooperative perception. *arXiv preprint arXiv:2403.16034*, 2024. 2, 3

[32] Yumin Xie, Langwen Zhang, Xiaoyuan Yu, and Wei Xie. Yolo-ms: Multispectral object detection via feature interaction and self-attention guided fusion. *IEEE Transactions on Cognitive and Developmental Systems*, 15(4):2132–2143, 2023. 8

[33] Xiaoqing Ye, Mao Shu, Hanyu Li, Yifeng Shi, Yingying Li, Guangjie Wang, Xiao Tan, and Errui Ding. Rope3d: The roadside perception dataset for autonomous driving and monocular 3d object detection task. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21309–21318, 2022. 1, 2, 3, 7

[34] Deng Yongqiang, Wang Dengjiang, Cao Gang, Ma Bing, Guan Xijia, Wang Yajun, Liu Jianchao, Fang Yanming, and Li Juanjuan. Baai-vanjee roadside dataset: Towards the connected automated vehicle highway technologies in challenging environments of china, 2021. 2, 3

[35] Haibao Yu, Yizhen Luo, Mao Shu, Yiyi Huo, Zebang Yang, Yifeng Shi, Zhenglong Guo, Hanyu Li, Xing Hu, Jirui Yuan, and Zaiqing Nie. Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21329–21338, 2022. 2, 3

[36] Heng Zhang, Elisa Fromont, Sébastien Lefevre, and Bruno Avignon. Multispectral fusion for object detection with cyclic fuse-and-refine blocks. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 276–280, 2020. 3

[37] Lu Zhang, Xiangyu Zhu, Xiangyu Chen, Xu Yang, Zhen Lei, and Zhiyong Liu. Weakly aligned cross-modal learning for multispectral pedestrian detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2, 3

[38] Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. Detrs beat yolos on real-time object detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16965–16974, 2024. 7

[39] Walter Zimmer, Christian Creß, Huu Tung Nguyen, and Alois C. Knoll. Tumtraf intersection dataset: All you need for urban 3d camera-lidar roadside perception. In *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, pages 1030–1037, 2023. 1, 2, 3, 7

[40] Walter Zimmer, Gerhard Arya Wardana, Suren Sritharan, Xingcheng Zhou, Rui Song, and Alois C. Knoll. Tumtraf v2x cooperative perception dataset. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22668–22677, 2024. 6